Nucleic Acids Research, 1994, Vol. 22, No. 9 1637–1639

Systematic sequencing of the *Escherichia coli* genome: analysis of the 2.4–4.1 min (110,917–193,643 bp) region

Nobuyuki Fujita, Hirotada Mori[1], Takashi Yura[1] and Akira Ishihama*
Department of Molecular Genetics, National Institute of Genetics, Mishima, Shizuoka 411 and
[1]Institute for Virus Research, Kyoto University, Sakyo-ku, Kyoto 606, Japan

Received January 24, 1994; Revised and Accepted March 28, 1994 DDBJ accession no. D26562

ABSTRACT

The complete sequence analysis of the *E.coli* genome was initiated as a collaborative study in Japan. Following the initial analysis of the 0–2.4 min region (Yura, T. *et al.* (1992) *Nucleic Acids Res.* 20, 3305–3308), a contiguous sequence of 82,727 bp corresponding to the 2.4–4.1 min region (110,917–193,643 bp as counted from 0 min) was determined. The resulting sequence was found to contain at least 33 known genes and 24 putative genes predicted from protein sequence homology.

INTRODUCTION

The knowledge of a complete DNA sequence of the genome from one cellular organism will provide us with a concept of the fundamental genetic framework essential for life. Two lines of large-scale sequencing of the *Escherichia coli* genome have emerged: one by the collaborative research group in Japan and the other by the Wisconsin group in USA. The first publication from the research group from Japan reported the sequences of 111,402 bp (1; DDBJ/EMBL/GenBank entry D10483). In this report, we describe the sequence of 82,727 bp (DDBJ/EMBL/GenBank accesion number: D26562) which lies directly clockwise to the previously sequenced region (1). A report from the third-step team from this country will be published in near future (Mizobuchi, K. *et al.*, in preparation).

MATERIALS AND METHODS

The following lambda phage clones from the Kohara's mini-set library were used for the sequencing: #111 (15B8), #113 (4E11), #114 (17C11), #115 (11C5), #116 (15A7) and #119 (21C8) (2). Nucleotide sequence was determined by combinations of the conventional methods as reported previously (1). Almost the entire sequence was determined for the clones #114 and #115, because only a few sequence data had been unequivocally mapped in this region. The complete sequence was also determined for the clone #119. The sequencing of total 71,095 bp was done by the research groups from three companies (including overlapping regions): Takara Shuzo, Co. (headed by M.Kitagawa), 10,585 bp; Toyo Jyozo, Co. (headed by J.Mizoguchi), 11,356 bp; and Fujiya, Co. (headed by M.Yamazaki), 49,154 bp, respectively. Computer analysis was done as reported previously (1). Protein similarity search was done against the PIR database release 35 (December, 1992). DNA similarity search was done against the DDBJ/EMBL/GenBank non-redundant DNA database maintained by DDBJ (DNA Data Bank of Japan) including daily update as of March 19, 1993.

RESULTS AND DISCUSSION

Construction of the contiguous nucleotide sequence

Newly determined sequences, (solid parts of the center bar in Fig. 1), were combined with the known sequences (solid arrows below the scale) in the DNA databases to construct a contiguous sequence of 82,727 bp. The 5′ end of this sequence overlaps by 485 bp with the 111,401-bp sequence reported previously (1) [one bp was erroneously counted in the published sequence], altogether accomplishing a contiguous sequence of 193,643 bp, covering the 0 to 4.1 minutes region of the *E.coli* genome.

Revision of the physical map

After the establishment of the ordered physical map of the whole *E.coli* chromosome (2), inconsistency or polymorphism has been noted at the 4 min region (3,4; Y.Kohara, personal communication). We then determined almost the entire sequence around this region, even though fragmentary sequences for this region have been published. As the result, it became evident that the phage 21C8 (#119) and 15A7 (#116) overlap by 1,566 bp, although these two clones were considered to be about 5 kb apart from each other in the revised Kohara's map (Y.Kohara, personal communication). Otherwise, the physical map predicted from the 82,727 bp sequence determined in this study agreed well with the published map (Fig. 2).

Identification of the putative genes

The 82,727 bp sequence contains 360 possible open reading frames consisting of 75 or more consecutive sense codons. The translated amino acid sequences from these ORFs were subjected to homology analysis against the PIR protein database. The

*To whom correspondence should be addressed

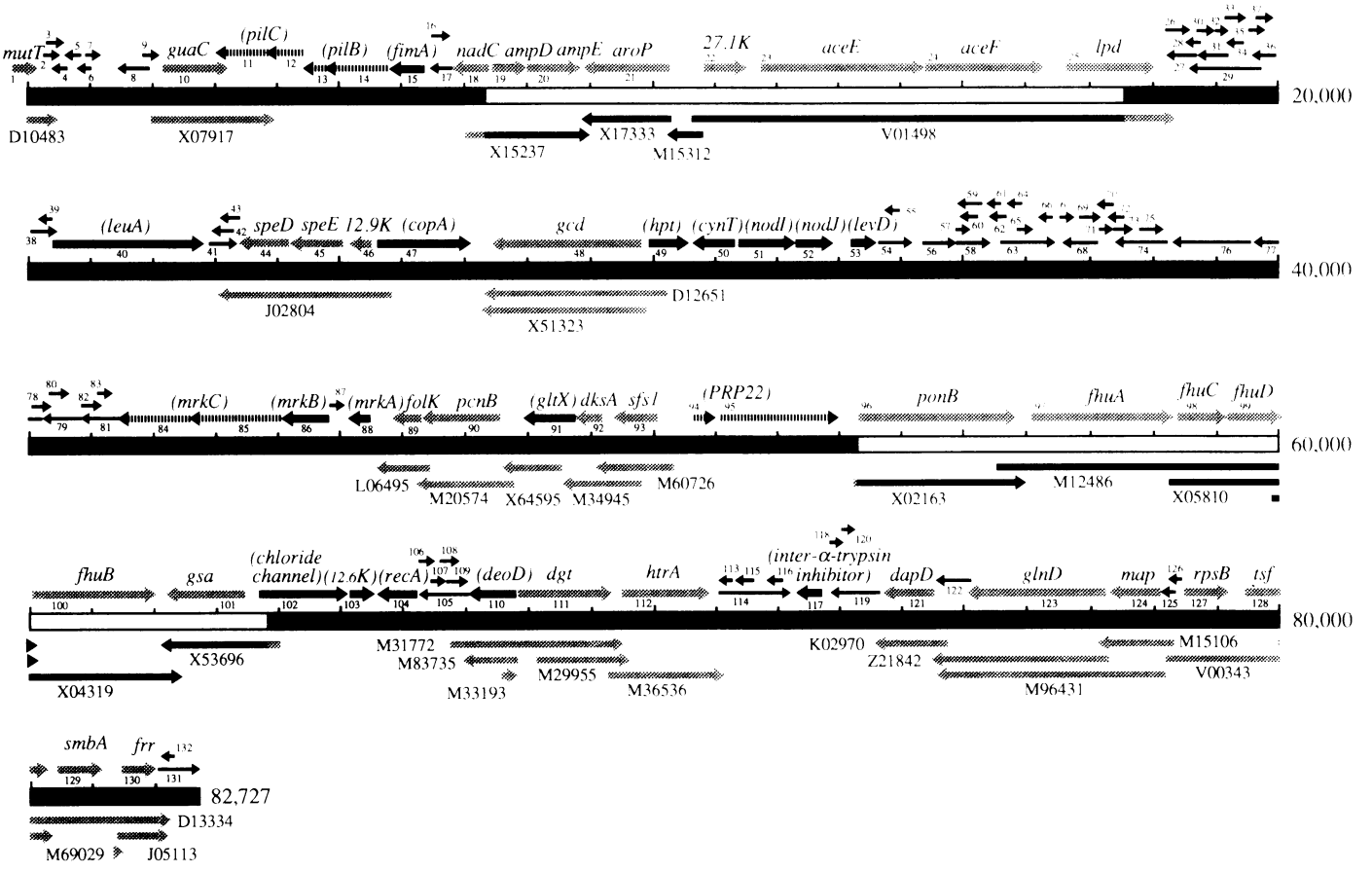

Figure 1. Possible ORFs predicted from the assembled sequence. Nucleotides are numbered from the first base of the EcoRI recognition sequence which lies in the middle of the *mutT* gene. Solid parts of the central bar indicate sequences determined in this study (see Materials and Methods). Solid arrows below the scale indicate data taken from the DDBJ/EMBL/GenBank databases (accession numbers are shown in the figure) to construct contiguous sequence. Other database entries which overlap with the assembled sequence are indicated by dotted arrows. All possible ORF's consisting of 75 or more consecutive sense codons were extracted and the predicted amino acid sequences were subjected to sequence homology analysis against the PIR database. Putative genes (thick solid arrows) predicted from the homology analysis and other uncharacterized ORF's (thin arrows) are shown along with known *E.coli* genes (thick dotted arrows). The lengths of the arrows represent the maximum number of codons (see Table 1) for putative genes and uncharacterized ORF's and the plausible or experimentally determined number of codons for known genes. Broken arrows indicate a group of neighboring ORF's whose amino acid sequences are similar to different parts of a single protein.

sequences of 33 ORFs agreed well with those of known *E.coli* proteins and the sequences of 24 ORFs showed considerable homology with those of known proteins from various organisms (see below). The majority of other ORFs exists within these protein sequences or otherwise overlap with these sequences (Fig. 1). Since overlapping genes are rare in the *E.coli* genes, these ORFs may not be translated into proteins. The non-overlapping ORFs with little or no homology to known proteins range from 15 to 20. Overall approximately 75 genes can be estimated to exist in the 2.4–4.1 min region of 82,727 bp in length (therefore, on average, one gene per 1.1 kbp DNA). Table 1 shows the list of open reading frames with significant degree of homology with one or more of the known proteins from various organisms. A number of ORFs carry sequences similar to those involved in biogenesis of bacterial pili and fimbriae. These are likely to be involved in the biogenesis of pili and fimbriae in *E.coli*. The genes coding for hypoxanthine phosphoribosyltransferase (*hpt*) and of purine-nucleotide phosphorylase (*deoD*) have not yet been determined in the *E.coli* chromosome. Thus, ORF49 and ORF110 may correspond to the *E.coli* genes for these enzymes. The ORF91 exhibited sequence similarity with the *E.coli gltX* gene encoding glutamate-tRNA ligase, suggesting that it may code for an as yet unidentified amino acid-tRNA ligase.

ACKNOWLEDGEMENTS

This work was supported by Grant-in-Aid for the Creative Basic Research (Human Genome Program; K.Matsubara, Osaka University) and by Grant-in-Aid for the Scientific Research on Priority Area (Comprehensive Anlaysis of the *E.coli* Genome; T.Yura, Kyoto University) from the Ministry of Education, Science and Culture of Japan.

REFERENCES

1. Yura, T., Mori, H., Nagai, H., Nagata, T., Ishihama, A., Fujita, N., Isono, K., Mizobuchi, K., and Nakata, A. (1992) *Nucleic Acids Res.* **20**, 3305–3308.
2. Kohara, Y., Akiyama, K., and Isono, K. (1987) *Cell*, **50**, 495–508.
3. Degryse, E. (1991) *Gene*, **102**, 141–142.
4. Heeswijk, W.V., Kuppinger, O., Merrick, M., and Kahn, D. (1992) *J. Bacteriol.* **174**, 1702–1703.

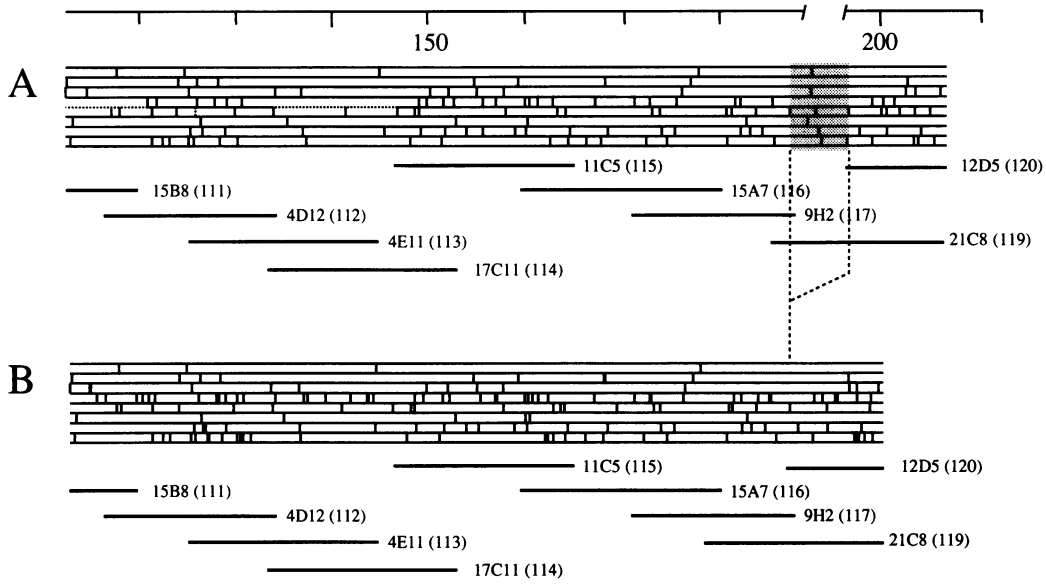

Figure 2. Revised physical map of the 2.4 to 4.1 minutes region of the *E.coli* genome. The revised version (Y.Kohara, personal communication) of the published physical map [**A**] was aligned with the map predicted from the assembled nucleotide sequence [**B**]. Each lane (from top to bottom) shows the restriction map for *Bam*HI, *Hin*dIII, *Eco*RI, *Eco*RV, *Bgl*I, *Kpn*I, *Pst*I, and *Pvu*II restriction enzymes, respectively. Horizontal bars below the map represent the chromosomal segments inserted in the respective lambda phage clones. The names of lambda clone (2) and the serial numbers in the mini-set library (in parentheses) are shown on the right side of the bars. Two physical maps can be best aligned if we assume that the original map contains an extra segment of unknown origin (about 6 kb in length, shown as a dotted area).

Table 1. Putative genes predicted from protein homology

ORF	Location[a]	Direction	No. of codons max.[b]	No. of codons plaus.[c]	Gene	PIR entry	Description	Score[d]	% match[e]
11:	3006 - 3932	L	309	276	(pilC)	B35384	pilC protein - Pseudomonas aeruginosa	281	26.7 (187)
12:	3848 - 4414	L	189	179	(pilC)	B35384	pilC protein - Pseudomonas aeruginosa	167	23.5 (153)
13:	4418 - 4966	L	183	139	(pilB)	A35384	pilB protein - Pseudomonas aeruginosa	343	42.9 (163)
14:	4792 - 5832	L	347	344	(pilB)	A35384	pilB protein - Pseudomonas aeruginosa	560	45.5 (224)
15:	5836 - 6369	L	178	134	(fimA)	S15266	fimA protein - Dichelobacter nodosus	175	45.9 (61)
40:	20404 - 22821	R	806	773	(leuA)	JQ0160	3-Isopropylmalate dehydratase - Mucor circinelloides	201	23.3 (343)
47:	25621 - 27114	R	498	474	(copA)	KSPSCY	Copper resistance protein precursor A - Pseudomonas syringae pv. tomato	210	24.9 (422)
49:	29944 - 30573	R	210	182	(hpt)	S10993	Hypoxanthine phosphoribosyltransferase - Vibrio harveyi	717	74.7 (174)
50:	30640 - 31305	L	222	213	(cynT)	QRECTC	Cyanate permease - Escherichia coli	205	32.5 (163)
51:	31381 - 32310	R	310	308	(nodI)	S13590	Nodulation protein nodI - Rhizobium meliloti	398	29.2 (305)
52:	32292 - 32918	R	209	203	(nodJ)	S08617	Nodulation protein nodJ - Rhizobium leguminosarum bv. trifolii	213	21.3 (197)
53:	33216 - 33620	R	135	104	(levD)	S11398	levD protein - Bacillus subtilis	128	25.8 (89)
84:	41431 - 42597	L	389	370	(mrkC)	D39142	mrkC protein precursor - Klebsiella pneumoniae	400	31.8 (333)
85:	42576 - 44042	L	489	443	(mrkC)	D39142	mrkC protein precursor - Klebsiella pneumoniae	801	38.2 (456)
86:	44062 - 44817	L	252	246	(mrkB)	C39142	mrkB protein precursor - Klebsiella pneumoniae	441	36.4 (236)
88:	45145 - 45489	L	115	112	(mrkA)	B39142	Type 3 fimbrial protein mrkA precursor - Klebsiella pneumoniae	110	31.8 (88)
91:	47941 - 48744	L	268	256	(gltX)	SYECET	Glutamate-tRNA ligase - Escherichia coli	364	35.5 (251)
94:	50647 - 51009	R	121	118		S13643	PRP22 protein - Saccharomyces cerevisiae	145	34.5 (119)
95:	51080 - 52999	R	640	613		S13643	PRP22 protein - Saccharomyces cerevisiae	423	32.3 (313)
102:	63715 - 65121	R	469	436		S16859	Voltage-gated choride channel protein - Pacific electric ray	194	25.6 (266)
103:	65170 - 65547	R	126	114		S04873	Hypothetical protein 118 (nifS 5' region) - Bradyrhizobium japonicum	210	37.1 (105)
104:	65600 - 66223	L	208	207	(recA)	S16386	recA protein - Vibrio cholerae	294	70.5 (78)
110:	67054 - 67797	L	248	232	(deoD)	A27854	Purine-nucleoside phosphorylase - Escherichia coli	127	21.4 (206)
117:	72308 - 72700	L	131	128		IYHU2	Inter-alpha-trypsin inhibitor complex component II precursor - Human	123	23.6 (127)

Location (a) of each ORF is shown on the basis of the maximum number (b) of contiguous sense codons bordered by two adjacent nonsense codons. Plausible number (c) of codons means the number of sense codons counted from the first ATG. When multiple proteins in the PIR database showed homology to a given ORF, they are represented by the one which showed the highest similarity score. The degree of protein similarity is presented by the 'optimized' FASTA similarity score calculated using the PAM250 matrix (d) and the percentage of identical amino acid residues in the homologous segment defined by FASTA (e). The number in parentheses shows the length of the homologous segment.